Functional Assessment of Erasure Coded Storage Archive

Josh Sackos
Washington State University
School of Electrical Engineering and Computer Science

Taylor Sanchez
California State University San Bernardino

Blair Crossman
New Mexico Tech
Science • Engineering • Research University

Mentors: HB Chen HPC-5 and Jeff Inman HPC-1

Abstract

As we approach exascale computing, disk storage becomes an attractive option due to scalability of disk bandwidth over tape drive. Disk failure rates make exascale archive systems prone to data loss. Replication is a possible solution, but can double or triple required storage space. Erasure coding is a promising option for an exascale archive because it offers the durability of replication with less overhead. The functionality of an erasure code archive system remains untested at exascale. Our project was to build and verify the functionality of two prototype erasure storage archives using commercial products from Scality and Caringo.

Both products had the functionality to read, write, balance, and rebuild data as well as offering metadata access. Caringo did not provide us with a POSIX gateway in time, but has a metadata indexing tool that allowed querying. We did not have the Scality indexing tool to query the metadata, but we were provided with the POSIX interface SFUSE. The POSIX gateway caused an average bandwidth loss of 70% for small files (< 1MB) and 50% loss for large files (> 1GB).

Overview

Object Storage

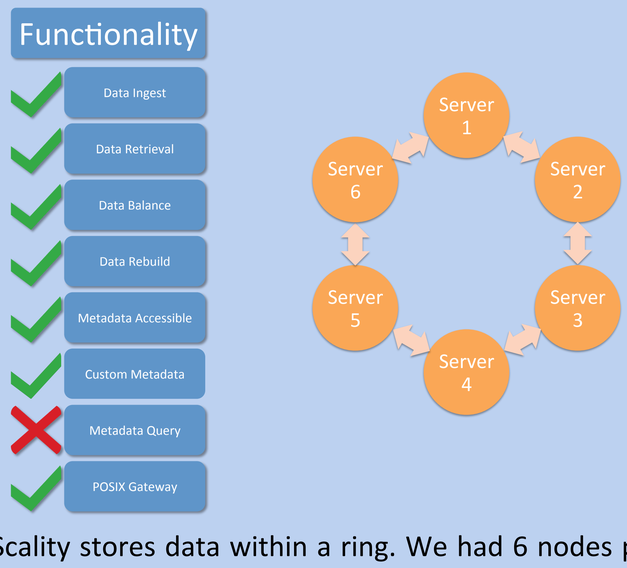
An object consists of data and metadata. Instead of fixed-size blocks, data objects are placed into flexible-sized data containers. The object storage paradigm scales with growing data demand.

Erasure Coding

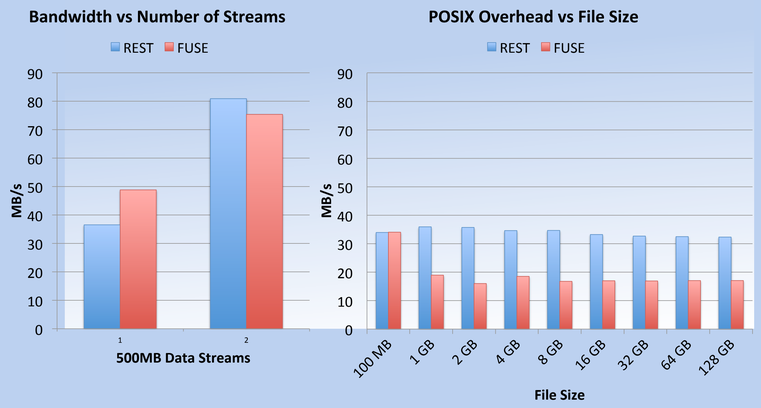
Offers data storage without centralized servers, and includes functionality to rebuild lost or corrupted data. The data repair capability offers large systems high durability and reduces hardware overhead. A file is broken into k data segments and n code segments and is dispersed throughout all servers in a system. Up to n data segments, code segments, or a mixture of data and code segments can be lost before data cannot be rebuilt, thus erasure coding has a resilience of n.

Scality

One Ring to Store Them All

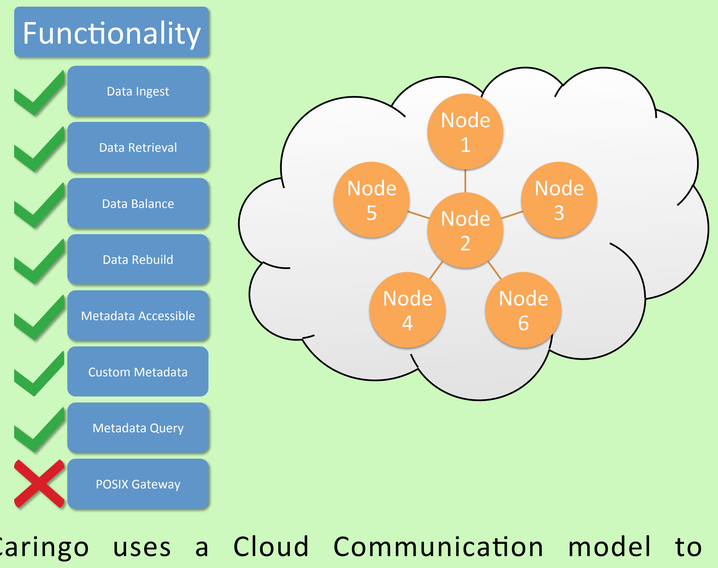


Scality stores data within a ring. We had 6 nodes per server, giving us a total of 36 nodes for our 6 servers. The 36 nodes form what is known as “The Ring”. Scality assigns key spaces to nodes to evenly distribute object storage locations. Scality met all testing requirements, except for metadata querying. This is due to the fact that we did not have their additional software ‘Mesa’ for indexing.



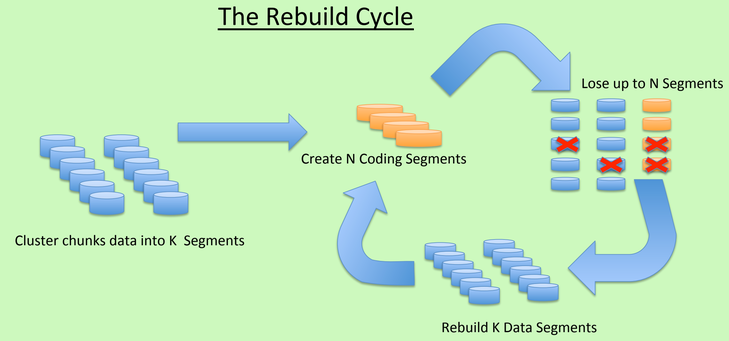
Caringo

Cloud Object Storage for a Rainy Day

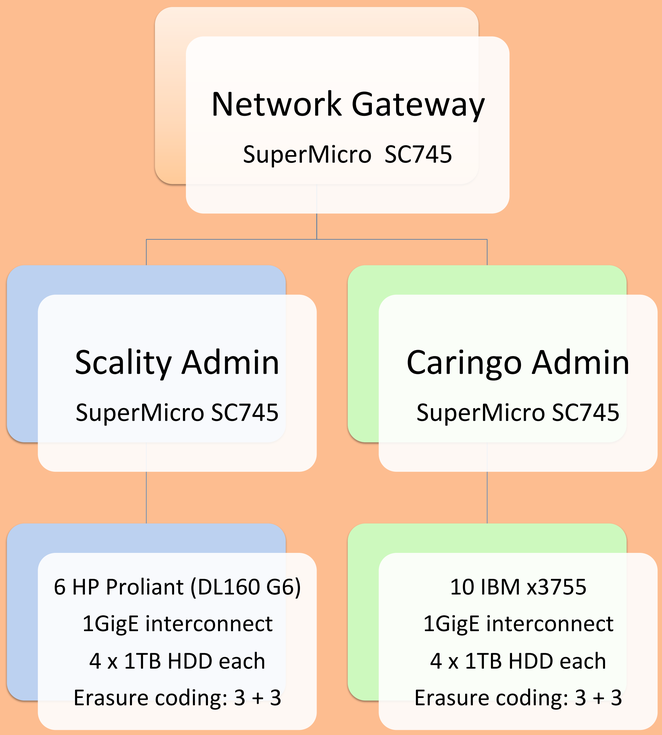


Caringo uses a Cloud Communication model to determine data storage. When a node is contacted with a new data stream, it sets up a "bid" where each of the nodes in the cluster sends in a bidding price for the stream. The lowest bid that is received is chosen and the stream is redirected to that node for storage. Caringo offers a service called the Content File Server that they advertise as a POSIX compliant gateway to the cluster. As this software was not available to us, Caringo failed the POSIX gateway functionality test.

The Rebuild Cycle



Testbed



Conclusions

We have found that the technology of erasure storage is a viable solution to LANL's archive system due to its scalability, parallelism, and robustness. Individually neither software meets the feature requirements. While Caringo has (what appears to be) a more mature solution, but did not provide a POSIX interface to us. Which was one of our requirements for the archive system. Scality has a higher potential to meet the needs of the lab as they are willing to work with their customers to engineer a more fitting solution. The two biggest downfalls of Scality were that they have only beta support for near unlimited data file size and currently require a stateful install (to a disk). Continued investment into REST interfaced erasure storage has great potential to find a replacement to tape drives backups.

Future Work

There is a need to continue testing if erasure storage can meet the speed requirements of high performance computing. There will need to be some calibration done, as the erasure storage systems are currently geared towards cloud (or internet) applications. There are already plans and designated clusters to continue testing and verify the viability of erasure storage.

Acknowledgements

Dane Gardner, Matthew Broomfield, HB Chen, Jeff Inman, Los Alamos National Laboratory, Probe, New Mexico Consortium, NSF, ISTI.

Contact Info

Taylor Sanchez: xTaylorSanchez@gmail.com
Josh Sackos: contact@joshsackos.com
Blair Crossman: Blair.Crossman@gmail.com

LA-UR-13-25968

Los Alamos National Laboratory · New Mexico Consortium · NSF · PRObE Parallel Reconfigurable Observational Environment · ISTI Information Science & Technology Institute